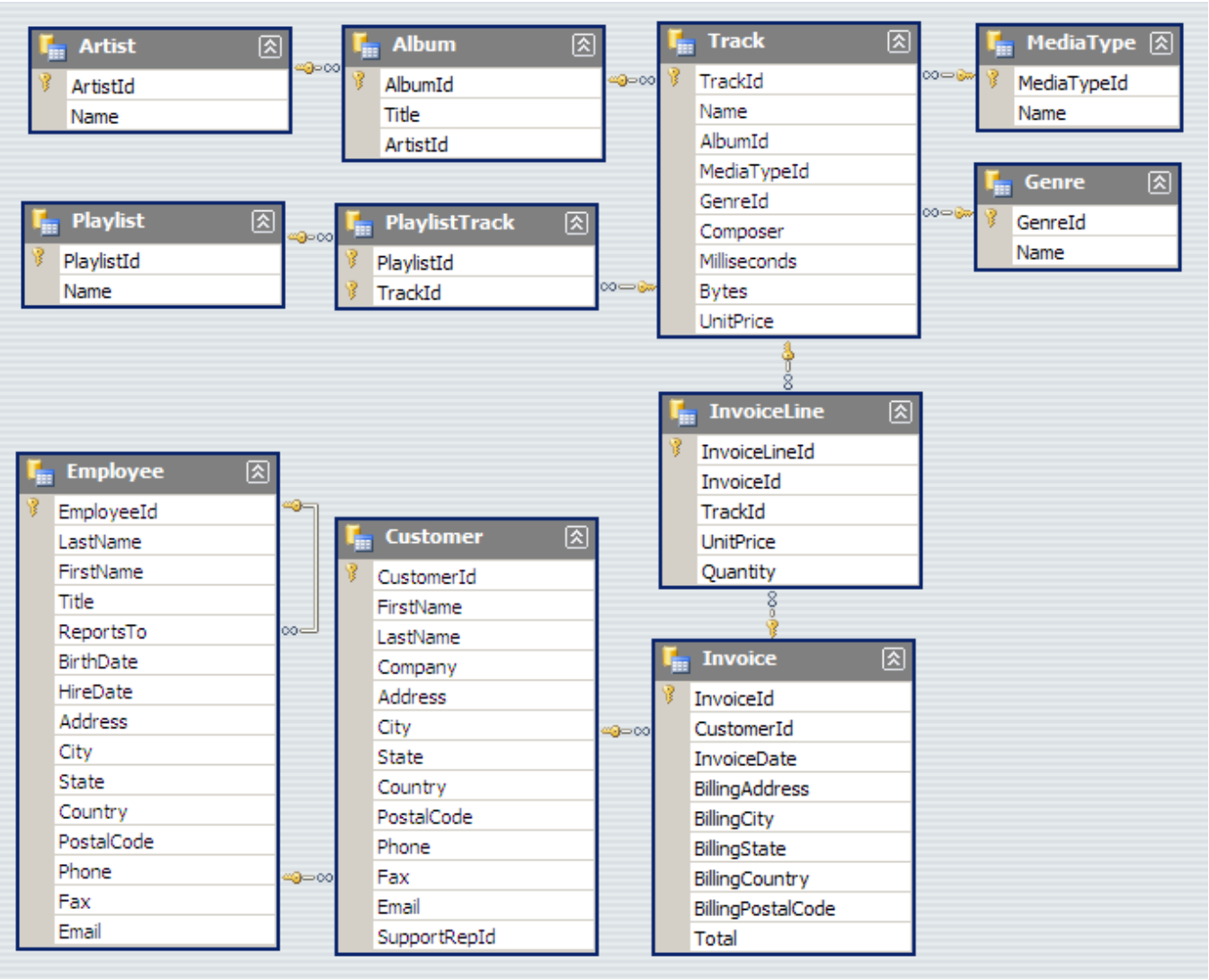
# Project Task Data warehousing / Business Intelligence

# The Chinook Sample Database

The Chinook Database is an OLTP system that is mainly used for keeping track of sales of an online music shop. Customers can buy music tracks. Tracks are part of an album which is in turn made by an artist.

Install the Chinook database from the backup file that is provided, **don't install it from the scripts you find on the internet because the data is different**!

# ERD Schema of the Chinook Database



# Introduction to PowerBI

Microsoft started in 2013-2014 developing new tools for reporting, having some unique characteristics like user friendliness, cloud integration and integration of all kinds of data sources. After a rather chaotic startup phase, in which several tools were introduced and withdrawn (e.g. PowerMap) it looks like at the time of this writing the dust has settled and the PowerBI environment has become stable.

PowerBI currently consists of three applications:

- Power BI Desktop: a desktop application to create reports and charts from all kinds of data sources (Excel, Facebook, Google Analytics, SQL Server, Oracle and many, many more) .
- Power BI Service: a cloud solution, integrated in Microsoft Azure where reports are published and viewed.
- Power BI Mobile: a mobile application for viewing reports.

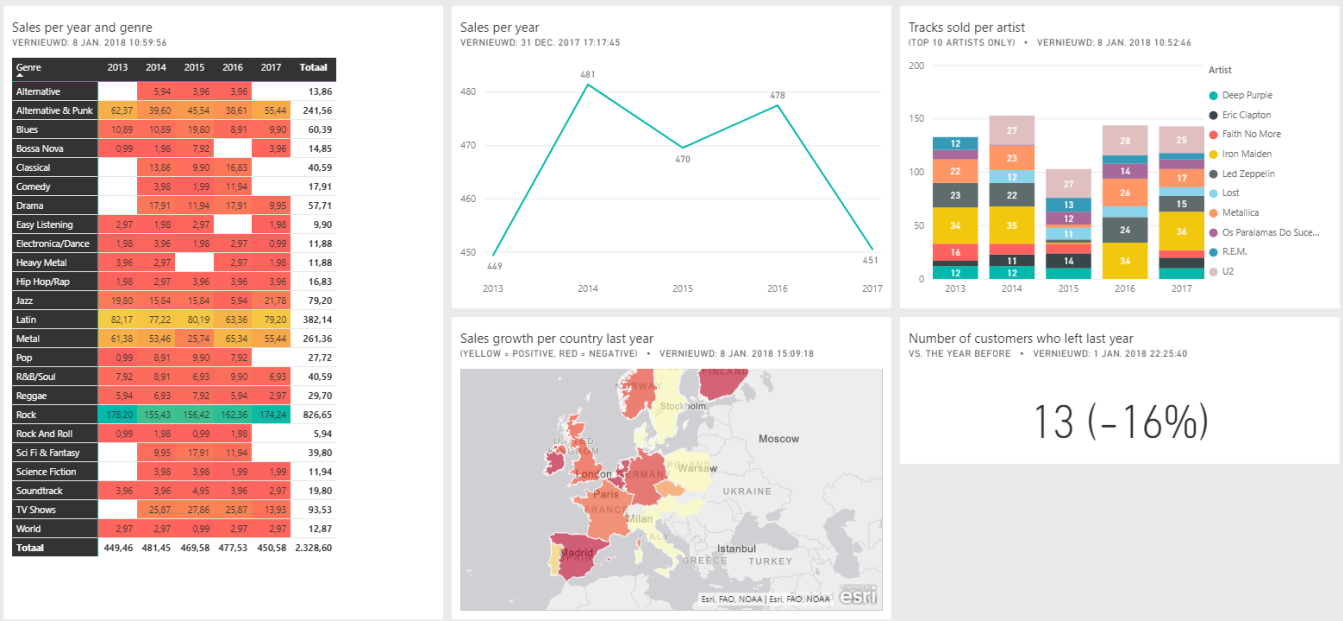
In this course we only use Power BI Service. You can also install the Power BI app to enjoy the reports you created on your smartphone.

Create a PowerBI account via "Use it free" on http://app.powerbi.com . Important remark: you need a professional e-mail address to be able to create an account. E-mail adresses @hotmail.com, @gmail.com, @telenet.be, etc. won't work. So USE YOUR HOGENT E-MAIL ADDRESS.

One way to get access to local files in Power BI is using OneDrive, the file share (say Dropbox like) solution of Microsoft. As a Office365 user you get 1TB online storage. If you are not an Office365 user you can also use OneDrive Personal.

# DASHBOARDING

A Dashboard is overview of graphics and statistics that show the current status of a business. It can be used for both following up operational and strategic goals. Below you see an example of a dashboard that is used by the company that runs the Chinook business. It is made by PowerBI service.

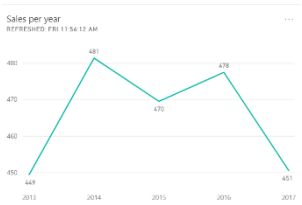



| Genre | 2013 | 2014 | 2015 | 2016 | 2017 | Totaal |
|---|---|---|---|---|---|---|
| Alternative | | 5,94 | 3,96 | 3,96 | | 13,86 |
| Alternative & Punk | 62,37 | 39,60 | 45,54 | 38,61 | 55,44 | 241,56 |
| Blues | 10,89 | 10,89 | 19,80 | 8,91 | 9,90 | 60,39 |
| Bossa Nova | 0,99 | 1,98 | 7,92 | | 3,96 | 14,85 |
| Classical | | 13,86 | 9,90 | 16,83 | | 40,59 |
| Comedy | | 3,96 | 1,99 | 11,94 | | 17,91 |
| Drama | | 17,91 | 11,94 | 17,91 | 9,95 | 57,71 |
| Easy Listening | 2,97 | 1,98 | 2,97 | | 1,98 | 9,90 |
| Electronica/Dance | 1,98 | 3,96 | 1,98 | 2,97 | 0,99 | 11,88 |
| Heavy Metal | 3,96 | 2,97 | | 2,97 | 1,98 | 11,88 |
| Hip Hop/Rap | 1,98 | 2,97 | 3,96 | 3,96 | 3,96 | 16,83 |
| Jazz | 19,80 | 15,84 | 15,84 | 5,94 | 21,78 | 79,20 |
| Latin | 82,17 | 77,22 | 80,19 | 63,36 | 79,20 | 382,14 |
| Metal | 61,38 | 53,46 | 25,74 | 65,34 | 55,44 | 261,36 |
| Pop | 0,99 | 8,91 | 9,90 | 7,92 | | 27,72 |
| R&B/Soul | 7,92 | 8,91 | 6,93 | 9,90 | 6,93 | 40,59 |
| Reggae | 5,94 | 6,93 | 7,92 | 5,94 | 2,97 | 29,70 |
| Rock | 178,20 | 155,43 | 156,42 | 162,36 | 174,24 | 826,65 |
| Rock And Roll | 0,99 | 1,98 | 0,99 | 1,98 | | 5,94 |
| Sci Fi & Fantasy | | 9,95 | 17,91 | 11,94 | | 39,80 |
| Science Fiction | | 3,98 | 3,98 | 1,99 | 1,99 | 11,94 |
| Soundtrack | 3,96 | 3,96 | 4,95 | 3,96 | 2,97 | 19,80 |
| TV Shows | | 25,87 | 27,86 | 25,87 | 13,93 | 93,53 |
| World | 2,97 | 2,97 | 0,99 | 2,97 | 2,97 | 12,87 |
| Totaal | 449,46 | 481,45 | 469,58 | 477,53 | 450,58 | 2.328,60 |

Behind every tile in the dashboard there is exactly one SQL statement.

# SQL exercises

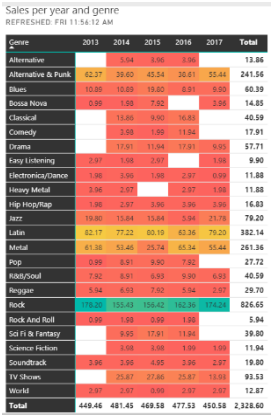
1. Write a query that shows the total sales per year
   a. Test your query in MS SQL Server Management Studio
   b. Create a view from the query. Name it "RPSalesPerYear"
   c. Save the query as a csv file on your OneDrive by right-clicking on the white square in the upper left corner. Name it RPSalesPerYear.csv. Add a title row to the file: "Year,Sales". Use your local settings separator, usually , or ; .

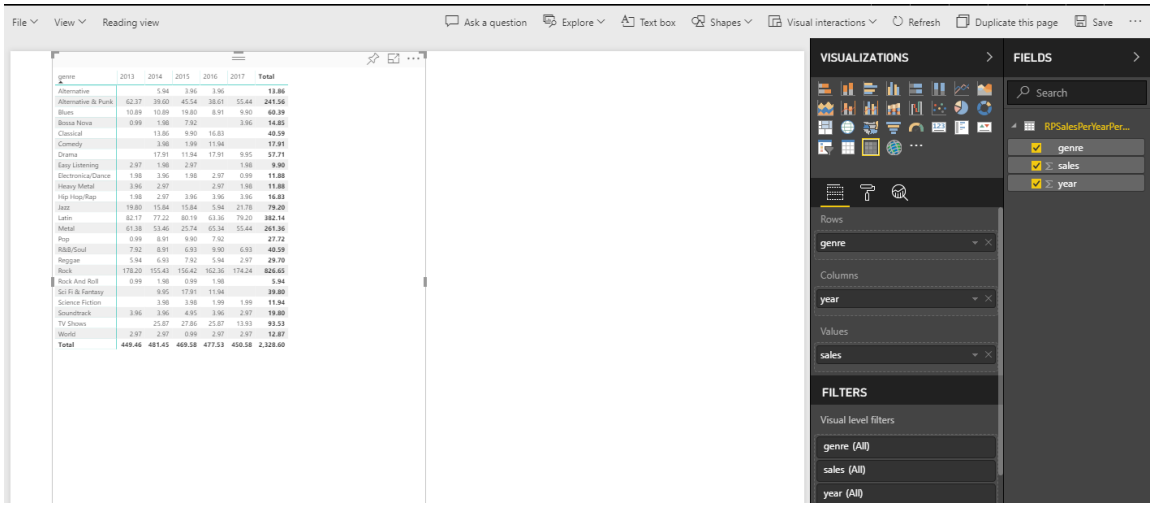



   d. Surf to app.powerbi.com and click "Create" in the upper right corner. Choose Report and click "import new data". Click "Files" and "OneDrive – Business". Explore the other options.
   e. Browse to your csv file and click "Connect". The data from your csv is now imported to the Azure platform.
   f. Click "View dataset". The report GUI is opened.
   g. Choose "Line chart" in the VISUALIZATIONS pane.
   h. Drag "year" to axis and "sales" to values. A line graph appears.
   i. Save your report under "File/Save" and Give it the name "RPSalesPerYear". Your report is now ready, but you can also add it to a dashboard.
   j. Go back to My Workspace and click "Create". Choose "Dashboard" and name it "Chinook".

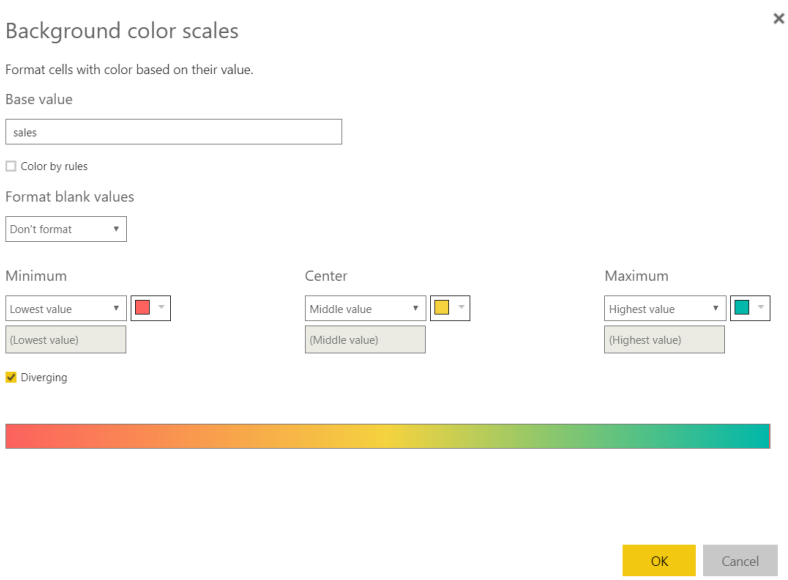
k. Go back to your report and click “Pin Visual”. This appears when hovering over the upper right corner above the report. The report is now added to the your dashboard.

2. Write a query that shows the total sales per year per genre
   a. Test your query in MS SQL Server Management Studio
   b. Create a view from the query. Name it “RPSalesPerYearPerGenre”
   c. Save the query as a csv file on your OneDrive. Name it RPSalesPerYearPerGenre .csv. Add a title row to the file: “year,genre,sales”. Use your local settings separator, usually , or ; .
   d. Surf to app.powerbi.com and click “Create” in the upper right corner. Choose Report and click “import new data”. Click “Files” and “OneDrive – Business”. Explore the other options.
   e. Browse to your csv file and click “Connect”. The data from your csv is now imported to the Azure platform.
   f. Click “View dataset”. The report GUI is opened.
   g. Choose “Matrix” in the VISUALIZATIONS pane.
   h. Drag genre to rows, year to columns and sales to values. A pivot table appears.



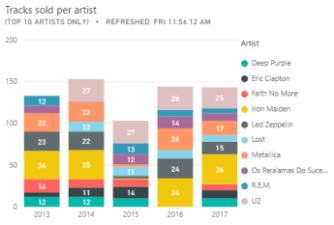

i. Click the format button in the VISUALIZATION pane.
j. Choose Matrix Style → Bold Header
k. Under “Field Formatting”, open “Conditional Formatting”. Set “Background color scale” to on, click “Advanced controls” and apply the setting below.




l. Save your report under "File/Save" and Give it the name "RPSalesPerYearAndGenre". Your report is now ready to add to your dashboard.
m. Click "Pin Visual" and position it on your dashboard.

3. We want to actualize the data.
   a. What's the year of the most recent invoice?
   b. Add a number of years to each invoicedate so the year of the most recent invoice is last year. E.g.: if the most recent invoice is currently of 2013 and this year is 2018, add exactly 4 years to each invoicedate BUT don't hard code "3". Instead, calculate this number and use it in your update statement. Working this way, it doesn't harm if you execute your update statement twice. Tip: have a look at the date functions in MS SQL Server.
   c. Select from the view you created in exercise 1 to check your update.
   d. Also export to csv file again and put the csv file on your OneDrive folder. Check that after approximately 30 minutes the update is reflected in the graph on PowerBI.

4. Write a query to show the number of tracks sold per artist per year for the top 10 best selling artists. *Take the overall top 10 for all years*.
   a. Test your query in MS SQL Server Management Studio
   b. Create a view from the query. Name it "RPTracksSoldPerArtistPerYearTop10"
   c. Save the query as a csv file on your OneDrive. Name it RPTracksSoldPerArtistPerYearTop10.csv. Add a title row to the file.
   d. Create a stacked bar chart for this data as shown in the upper right corner of the dashboard above.
   e. Add this chart to your dashboard.



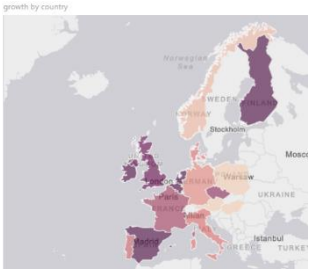

5. Create a new table COUNTRY by importing the csv file country.csv.
   a. Right click on the name of the database, choose "Tasks" / "Import Data" and follow the wizard. Choose "Flat file" as Source and "SQL Server Native Client" as Destination.
   The first lines of country.csv look like this:

```
name,alpha-2,alpha-3,country-code,iso_3166-2,region,sub-region,region-code,sub-region-code
Afghanistan,AF,AFG,4,ISO 3166-2:AF,Asia,Southern Asia,142,34
```

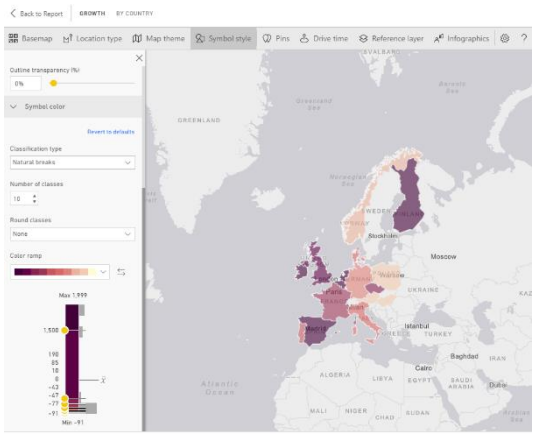
```
Åland Islands,AX,ALA,248,ISO 3166-2:AX,Europe,Northern Europe,150,154
Albania,AL,ALB,8,ISO 3166-2:AL,Europe,Southern Europe,150,39
Algeria,DZ,DZA,12,ISO 3166-2:DZ,Africa,Northern Africa,2,15
American Samoa,AS,ASM,16,ISO 3166-2:AS,Oceania,Polynesia,9,61
Andorra,AD,AND,20,ISO 3166-2:AD,Europe,Southern Europe,150,39
Angola,AO,AGO,24,ISO 3166-2:AO,Africa,Middle Africa,2,17
Anguilla,AI,AIA,660,ISO 3166-2:AI,Americas,Caribbean,19,29
```

b. Ensure in table customer only countries that exist in table country can be entered. Does this work? Why not?
c. Solve this problem by renaming the country in the table countries.

6. Show last year's sales growth per country for all <u>European</u> countries. More specific, show on a map of Europe for each country a colour that reflects last year's growth as compared to the year before. Express the growth as a percentage. To avoid division by zero, hardcode 1999 % if there were no sales two years ago.



   a. Import the file country.csv. Right-click on "Chinook", Choose Tasks/Import Data. Choose "Flat file source" as the source and "SQL Server Native Client" as the destination. Leave the other settings unchanged.
   b. Write a query to calculate the growth for each European country.
   c. Save the result set as a csv and create a dataset in PowerBI as above.
   d. Choose "ArcGIS Maps for PowerBI" in the VISUALIZATIONS pane.
   e. Drag country to location and growth to color.
   f. Click on the ellipsis (…) in the upper right corner of your report and choose "Edit".
   g. Select the colours of your choice.
   h. Publish the report on your dashboard.



# DATAWAREHOUSING

## Design the data warehouse

7. ***Design a data warehouse according to the star schema that allows to create in a straightforward way high performant reports.*** Some of the reports needed are:
    - Evolution of the monthly sales in terms of total billing amount, number of tracks sold and total milliseconds sold.
    - Evolution of the monthly sales per genre.
    - Several ad-hoc reports based on customer data (city and country), genre and artist.

    Add the appropriate primary and foreign keys.

## Fill the data warehouse

8. Create, using Microsoft Visual Studio (SSDT), a package to fill the facts and dimensions. Make sure your package can be repeatedly run. **Consider** (only) **DimCustomer as a slowly changing dimension.**

    Add test data (facts and dimensions) between consecutive runs of your package and check the results.

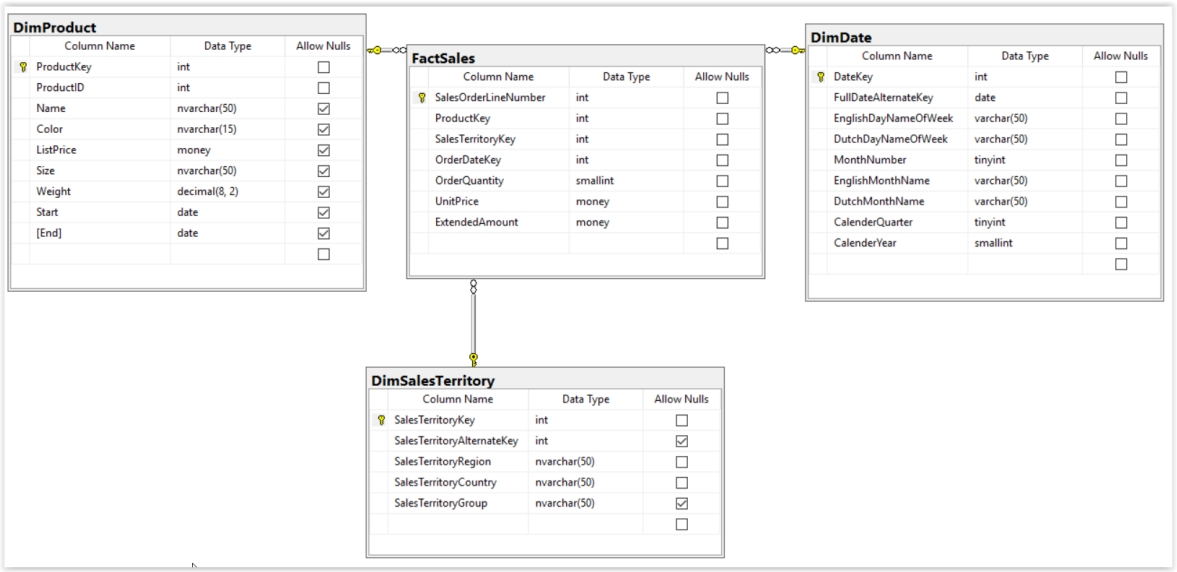
    Schedule the package every night at 2am.

## Query the data warehouse

9. Rewrite the queries 1, 2, 4 and 6 on the data warehouse. Save each query as a view.
10. Create a package in SSDT to export the results of each query to csv file on your OneDrive folder.
11. Schedule the package every hour.
12. Create a report in PowerBI service report for each csv as above and create a new dashboard that shows all results.

## Deliverables

1. Screen dumps of both dashboards (exercises 1-6 and 12)
2. View definitions (= queries) for exercises 1-6 and 9
3. Screenshot ERD diagram of dataware house, showing data types. E.g.:




4. ETL
    a. Screen dump of your control flowchart
    b. Screen dump of your data flow for the slowly changing dimension
    c. Zip file of the complete Visual Studio Solution directory.